ECBD: Evidence-Centered Benchmark Design for NLP

Anonymous ACL submission

Abstract

Benchmarking is seen as critical to assessing progress in NLP. However, creating a benchmark involves many design decisions (e.g., which datasets to include, which metrics to use) that often rely on tacit, untested assumptions about what the benchmark is actually measuring. There is currently no principled way of analyzing these decisions and how they impact the validity of the benchmark's measurements. To address this gap, we draw on evidence-centered design in educational assessments to propose `ECBD` (Evidence-Centered Benchmark Design). Our framework formalizes the benchmark design process into five modules and specifies the roles of each module and their interplay in collecting the evidence necessary to support the benchmark's measurement. We demonstrate the use of `ECBD` by conducting case studies with three benchmarks: BoolQ, SuperGLUE, and HELM. Our analysis reveals common trends in benchmark design and documentation that could threaten the validity of benchmarks' measurements.

1 Introduction

Benchmarking has long been seen as critical to assessing the progress of natural language processing (NLP) models and guiding their selection for downstream applications. As zero-shot and in-context learning with language models (LMs) have become prevalent, evaluation in NLP has shifted from measuring model performance on a specific dataset to using large benchmarks that cover multiple linguistic tasks (e.g., GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), BIG-Bench (Srivastava et al., 2022), HELM (Liang et al., 2022), etc.). These benchmarks are growing larger and more ambitious (e.g., HELM aims to "assess language models in their totality"), covering an ever-increasing number of measured capabilities with ever-increasing numbers of datasets and metrics.

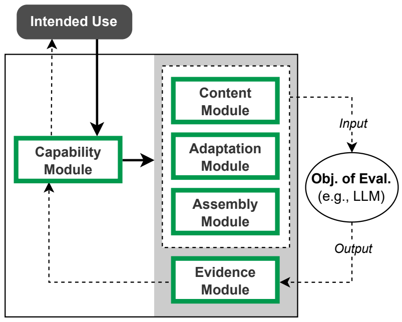

Figure 1: Simplified schema of the Evidence-Centered Benchmark Design (`ECBD`) framework. Solid line arrows indicate the process of designing a benchmark (e.g., designers should decide on the intended uses of the benchmark before deciding what capabilities are of interest). The dotted line arrows indicate the process wherein the benchmark gathers necessary evidence.

This trend increases the complexity of assessing the quality of a benchmark. Do benchmark results—most often in the form of numerical scores—provide meaningful insights about the evaluated models? For what purposes are these results useful? Are the benchmark measurements valid? The field of NLP lacks a systematic way of reflecting on these important questions. Such issues with test validity do not only concern NLP benchmark designers. In parallel, researchers and practitioners in educational testing often face similar questions: do students' exam results provide meaningful insights about their ability in, for example, reading comprehension? Can these results be used to determine whether a student needs remedial classes?

In this work, we take inspiration from the Evidence-Centered Design (ECD) framework in educational testing—which guides the process of creating, documenting, and validating tests—and propose Evidence-Centered Benchmark Design (`ECBD`) framework, in which we view benchmarking as the process of gathering evidence from objects of evaluation (e.g., language models) about whether or to what degree they have some capabilities of interest.

ECBD unpacks and formalizes benchmark design decisions into five modules, each having a specific role in supporting the process of collecting necessary evidence (see Figure 1). For each module, we provide guiding questions that help benchmark designers document, justify, and validate their design choices. These same questions also guide the analysis of existing benchmarks: what are the design decisions shaping the benchmark, why did its creators make these decisions, and what evidence do they provide to support their decisions?

To illustrate the usage of this framework in benchmark analysis, we apply it to three different benchmarks: BoolQ (Clark et al., 2019), SuperGLUE (Wang et al., 2019), and HELM (Liang et al., 2022). ECBD allowed us to find common practices, such as poor conceptualization of capabilities, that threaten the validity of these benchmarks' measurements. In general, we find that these benchmarks lack justification and validation.

2 Background & Related Work

Benchmarking in NLP At a time when most NLP systems were built for a single specific task, Wang et al. (2018) introduced the benchmark, General Language Understanding Evaluation (GLUE), with the goal of helping the research community develop models with more *general* language understanding ability. It is a collection of nine English sentence understanding tasks, covering question answering, sentiment analysis, and textual entailment. In around a year, the performance of evaluated LMs surpassed that of non-expert humans on the benchmark, prompting the development of SuperGLUE (Wang et al., 2019), whose main contribution is the increased difficulty of included tasks.

This trend of evaluating models across an increasing number of datasets continues with recent benchmarks such as XTREME (Hu et al., 2020), covering 40 languages, and GEM (Gehrmann et al., 2021, 2022), covering language generation tasks. Collaborative benchmarks such as BIG-Bench (Srivastava et al., 2022), now counting more than 200 tasks in its repository,[1] encourage the research community to add on new tasks.

Our proposed framework encourages a critical analysis of these increasingly complex benchmarks and guides reflection surrounding their validity.

Critiques and Meta-Analyses Much prior work has surveyed and critiqued NLP evaluation and machine learning (ML) evaluation in general. Bowman and Dahl (2021) outline a list of criteria that useful benchmarks for natural language understanding (NLU) should meet, including validity. Similarly, Wagstaff (2012) highlights the disconnect between benchmark results and real world impact for ML evaluation—does a given increase on the benchmark actually lead to positive impact in the tested domain of application?—while Liao and Xiao (2023) argue for centering large language model evaluation on how models will be used in practice. Analyses of benchmarks in NLP evaluation have raised concerns about annotation artifacts (Gururangan et al., 2018), threats to validity (Blodgett et al., 2021), lack of justification surrounding design choices (Goldfarb-Tarrant et al., 2023), inconsistent results from benchmarks aimed at measuring similar things (Akyürek et al., 2022), and benchmarks' lack of robustness (Alzahrani et al., 2024).

Documentation in NLP and Machine Learning Various documentation guidelines have been proposed across NLP and machine learning. Datasheets for Datasets (Gebru et al., 2021) provides a standardized process for documenting machine learning datasets, formulated as a list of questions (e.g., "Does the dataset contain data that might be considered confidential?"). In NLP, Data Statements for NLP (Bender and Friedman, 2018) contains guidelines more specific to speech and text data, asking practitioners to document details about how data is curated such as the demographics of the speakers included, while model cards (Mitchell et al., 2019) have been proposed to document model characteristics.

Our work contributes a set of guidelines for documenting NLP benchmark choices, with a particular focus on choices that build the process of gathering necessary evidence about whether, or to what degree, an evaluated model has the capabilities of interest. For example, in guiding data documentation, our framework differs from prior work by focusing on how this data is used in the benchmark to produce measurement. Beyond guiding documentation, our proposed framework also guides the process of validating the benchmark.

Measurement Theory In the social sciences, hypothesized theoretical entities known as **constructs** (e.g., a person's creativity, attitude towards a social issue) cannot be directly measured. Instead, the measurement is indirect, relying on samples of

[1] https://github.com/google/BIG-bench

observable behaviors obtained through **tests**. Measurement theory is the study of test development, aiming to minimize measurement error so to produce the best measures of the desired constructs (Bandalos, 2018). Educational testing is rooted in measurement theory, aiming to produce the best measures of students' abilities.

The quality of tests depends on their **validity**, which refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association, 2014). Bandalos (2018) argues that it is the most important quality of a test as it concerns the fundamental issue of what measurement instruments (i.e., tests) are really measuring.

These concepts are relevant to NLP, as desirable model capabilities (e.g., language understanding) most often cannot be directly measured; they are unobservable constructs, and NLP benchmarks can be seen as tests that use observable model behaviours (e.g., LM-generated text) to measure these constructs. Thus, the validity of NLP benchmarks is also a critical concern (Bowman and Dahl, 2021; Blodgett et al., 2021; Fleisig et al., 2023).

Evidence-Centered Design (ECD) in Education is a framework introduced in the field of education with the goal of guiding the design, evaluation, and interpretation of educational tests (Mislevy, 2003). Our main source of inspiration to create Evidence-Centered Benchmark Design (`ECBD`) comes from the conceptual assessment framework (CAF), a vital component of ECD consisting of five models:

i) **Student model:** specifies the constructs that characterize the students and that the test aims to measure. This model connects the test to its intended uses (e.g., if a test is to determine whether students need remedial language classes, should their reading comprehension skill be measured?).

ii) **Task model:** builds a pool of tasks (i.e., test items) that draw out responses from students. Since the test relies on these responses to measure the constructs of interest, the tasks should elicit evidence about said constructs.

iii) **Presentation model:** specifies how a given test item is presented to students (e.g., font size, instructions given by teachers). The goal is to avoid introducing measurement error—e.g., due to differences in the readability of the test due to font size.

iv) **Assembly model:** specifies how tasks are selected from the available pool to be presented to students (e.g., when there are 100 exam questions but students can only answer 20, how should the test select these 20 questions?). This model monitors the amount of evidence that will be collected (e.g., are the selected 20 questions sufficient to measure reading comprehension?)

v) **Evidence model:** specifies how to measure constructs specified in the student model by observing students' performance on the presented test items. It consists of two components: one specifies item-level scoring (i.e., extracting evidence from students' performance on a single test item) and the other specifies test-level scoring (i.e., accumulating extracted evidence across all presented test items).

In summary, each CAF model has specific roles to fulfill, and together they roadmap the process of educational testing. We adapt CAF models to NLP benchmarking, proposing a framework for benchmark design that similarly centers evidence in measurement.

3 Evidence-Centered Benchmark Design

We consider benchmarking as the process of gathering, from objects of evaluation (e.g., LMs), capability evidence—i.e., evidence about whether or to what degree said objects have the capabilities of interest. Evidence-Centered Benchmark Design (`ECBD`) structures this process into five modules,[2] each of which has a specific role in the collecting necessary capability evidence: the capability module (§ 3.1), the content module (§ 3.2), the adaptation module (§ 3.3), the assembly module (§ 3.4), and the evidence module (§ 3.5).

In addition, for each module, `ECBD` decomposes the design process into three actions. To guide benchmark creation, `ECBD` requires benchmark creators to i) **describe** their design choices; ii) **justify** them, forming a hypothesis about how these choices ensure that the module accomplishes its role; and iii) further **support** these hypotheses, which requires gathering another type of evidence—*validity evidence*.[3] Such evidence can be theoret-

[2]In adapting ECD, we have renamed some terms to avoid confusion: i) *module* instead of CAF *model*, as *model* often designates an *NLP model*; ii) *content* instead of *task*, as *task* often refers to a category in the context of NLP (e.g., the task of question answering) instead of a single instance (e.g., a single exam question).

[3]For clarity, capability evidence is about the capabilities of interest, and is gathered by the benchmark about the object of

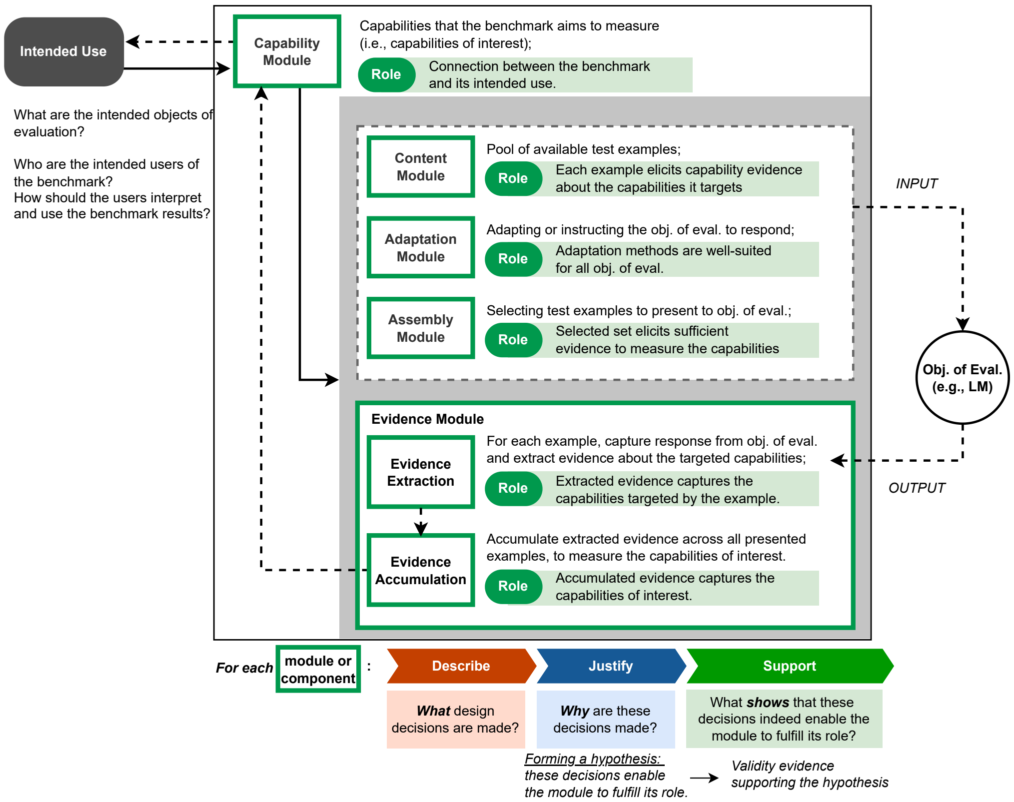

Figure 2: The Evidence-Centered Benchmark Design framework. Solid line arrows indicate the process of designing a benchmark (e.g., designers decide on the intended uses of the benchmark before deciding what capabilities are of interest). The dotted line arrows indicate the process of the benchmark gathering necessary capability evidence.

ical (e.g., theoretical work supporting the definition of a capability) or empirical (e.g., experiments correlating metric scores with some ground-truth scores). In addition to helping benchmark creators reflect on their design choices, ECBD helps benchmark analysis—e.g., performed by benchmark users or third parties—by drawing attention to whether and how benchmark creators describe and justify their design decisions, and to what extent there is validity evidence supporting these decisions. We illustrate our proposed framework in Figure 2, and to facilitate its usage, we formulate it as a worksheet of 20 questions (Appendix A).

Intended Use While clearly establishing the intended use of a benchmark is not a ECBD module, it is a step that must precede benchmark design or analysis using ECBD modules. This first step is crucial because the validity of a benchmark concerns whether it can be used as intended. The framework asks: i) What are the intended objects of evaluation (analogously, "test takers")? ii) Who are the intended users of the benchmark, and how should they interpret and use the benchmark results? If the intended use is not clearly stated at first, designers risk making choices simply because they are convenient or common practices, likely resulting in a benchmark that does not serve any particular purpose. Furthermore, if the the intended use is not clearly communicated to potential users of the benchmark, they could unintentionally misuse it (e.g., use it to evaluate other objects than the intended ones), or misinterpret its results.

3.1 Capability Module

The capability module specifies the capabilities—constructs that the objects of evaluation are thought to exhibit or possess—that the benchmark aims to measure. This module should be ***the connection between the benchmark and its intended use:*** what capabilities to measure depends on the benchmark's intended use. This module requires benchmark designers to define the capabilities of interest, justify the aforementioned connection, and gather validity evidence supporting the choices and definitions of the capabilities. This process encourages reflection on i) the definitions, which are often contested and may depend on context (e.g., users of the evaluated NLP systems and their needs), and ii)

evaluation. Validity evidence is about the benchmark design choices, gathered by the benchmark creators themselves or other parties (e.g., other NLP researchers, benchmark users).

the relevance of measured capabilities to the benchmark's intended use, as measuring irrelevant capabilities or overlooking relevant ones could threaten the validity of the benchmark.

Survey studies on relevance of capabilities could provide validity evidence for this module. Liao et al. (2022) is an example of such a survey study for explainable AI algorithms, where topical experts and end-users were asked what evaluation criteria are of importance for such algorithms.

3.2 Content Module

The content module specifies the pool of available test examples that the benchmark could require objects of evaluation to perform or to respond to. These examples should ***elicit evidence about the capabilities of interest***, so that this evidence can be later extracted from the responses and accumulated to produce measurements of those capabilities. Note that it is not necessary for each example to target *all* capabilities of interest, as examples can be used in combination (see Section 3.4).

Through the characteristics of the test examples, benchmark designers should be able to justify how each example elicits evidence about the capabilities it targets. Gathering validity evidence for this module could involve analysis by experts who assess whether test examples capture the capabilities of interest.[4] The study by Blodgett et al. (2021) is an example of such an analysis for NLP benchmarks measuring stereotyping. They identify, for instance, test examples that contain true facts instead of harmful stereotypes (e.g., "Afghanistan shares a border with Pakistan. Most people there are Muslim." (Nangia et al., 2020)). An evaluated model favoring such examples is likely not indicative of the model having harmful biases. Consequently, the prevalence of such test examples threatens the validity of these benchmarks.

3.3 Adaptation Module

When evaluating models or systems, benchmarks might employ a myriad of methods that i) adapt the models/systems (e.g., fine-tuning), or ii) format or add onto the test example (e.g., adding non-test examples in few-shot prompting). These methods, specified in the adaptation module, should be chosen carefully so as to not confound benchmark results: ***they should be well-suited to all objects of evaluation and not disadvantage some objects***.

[4]In measurement theory, this type of validity evidence is referred to as "content validity."

For example, if a benchmark employs prompting for LMs, some LMs might respond poorly to certain prompt formats, thus confounding benchmark results; poor performance might be indicative of this sensitivity to prompt formatting instead of providing meaningful information about the capabilities of interest.

3.4 Assembly Module

The pool of available examples specified by the content module (Section 3.2) is what the benchmark has available to use. The assembly module specifies which examples from this pool are actually used by the benchmark for evaluation, and whether this subset ***allows the benchmark to gather sufficient evidence for all capabilities of interest***.

The simplest assembly method would be to use all available examples. When there are resource constraints (e.g., computational resources, financial resources, or time), it may become necessary to consider more sophisticated assembly methods to preserve the quality of the benchmark—i.e., using fewer test examples should not introduce an unacceptable amount of measurement error.

3.5 Evidence Module

The evidence module specifies how capability evidence is extracted from responses obtained from objects of evaluation (evidence extraction), and how this evidence is accumulated to produce benchmark results that measure the capabilities of interest (evidence accumulation).

Evidence Extraction For each presented test example, objects of evaluation produce observable responses (e.g., LM-generated text, token probabilities). Evidence extraction involves specifying what responses are captured by the benchmark and how the benchmark ***extracts evidence, from these responses, about the capabilities targeted by the test example***.

This process necessarily involves representing the evidence, which is still an abstract concept at this point, via some observable variables such as numerical scores (e.g., 1/0 to indicate that a LM-generated text is fluent/disfluent, representing a piece of evidence about the LM's ability to generate fluent text). So benchmark designers need to justify and show that these variables actually capture the target capabilities. For example, experiments examining the relationship (e.g., correlation) between automatic metric scores and human-annotated scores (assumed to be ground-truth) could provide empirical evidence for this component.

Evidence Accumulation Benchmarks involving multiple test examples need to accumulate multiple pieces of extracted evidence to produce the measurement of the capabilities of interest—the benchmark results to be interpreted and used. This component thus connects observable variables from evidence extraction to the capability module (Section 3.1): ***the accumulated evidence should capture the capabilities of interest***. For example, the results of a benchmark could be the average of example-level scores if the distribution of example-level scores is assumed to follow a normal distribution. Gathering validity evidence could involve testing this assumption about the distribution.

4 Case Studies

To illustrate how our framework guides benchmark analysis and helps foreground possible validity concerns, we apply the ECBD worksheet to the analysis of HELM (Liang et al., 2022), SuperGLUE (Wang et al., 2019), and BoolQ (Clark et al., 2019).

4.1 Analyzed Benchmarks

SuperGLUE aims to be *"a more rigorous test of language understanding"* than its predecessor GLUE (Wang et al., 2018). It includes 8 pre-existing datasets, each corresponding to a *"language understanding task."* HELM, the most recent benchmark of the three, is meant to be a *"living benchmark"* to be continuously updated. When its accompanying paper was first published, HELM included 15 existing datasets.[5] BoolQ, which is re-used in both SuperGLUE and HELM, includes a novel dataset of naturally occurring yes/no questions.

These benchmarks are different in many ways: they are from different points in time, are of various sizes, aim to capture different capabilities, and are built differently (e.g., BoolQ being a novel dataset versus SuperGLUE and HELM re-using existing datasets). Due to its' flexibility, ECBD can be applied to all these benchmarks.

[5]HELM includes two evaluations that seem to be completely independent: a "core" evaluation and a supplementary "targeted" evaluation. As the main focus of the accompanying paper is on the former, we consider it as a single, independent benchmark that we focus on for our analysis.

4.2 Method

The ECBD worksheet for each benchmark is completed by two to three authors of this paper, where one author first read the paper introducing that benchmark, and then re-read it while completing the worksheet. One to two other authors then examined the completed worksheets while reading the paper. We discussed and resolved any ambiguities and uncertainties that arose during this process. The completed worksheets can be found in the Supplemental Material.

4.3 Observations

We overview key concerns with the design of existing benchmarks that ECBD's modules help us foreground.

Intended use: Benchmarks' intended uses are vaguely specified. Specifying a benchmark's intended uses is a crucial first step in ECBD. By examining how the three benchmark discuss their intended uses, we found little description of who their intended users are. In particular, we found no explicit mentions of intended users for BoolQ and SuperGLUE. Across all three benchmarks, it is also unclear how benchmark users should interpret and use the benchmark results, with HELM explicitly stating that the use and interpretation of benchmark results is up to the users to decide for themselves.[6] Since validity involves whether the benchmark results can be used as intended, this lack of information makes the analysis and validation of these benchmark difficult. In particular, it is difficult to assess whether measured capabilities are relevant to the intended use of the benchmarks.

Capability module: When evaluating complex capabilities, benchmarks seem to break down capabilities of interest into sub-capabilities that are perhaps easier to measure, but this process is sometimes not explicitly described. ECBD's capability module draws attention towards what capabilities the benchmarks measure and how they are conceptualized. For SuperGLUE, which aims to measure "general language understanding" (GLU), we found that the benchmark seems to consider intermediate capabilities of interest that contribute

[6]*"[W]e expect the totality of the results we provide are not relevant for every practical use case: we anticipate practitioners should first identify scenarios and metrics pertinent to their use conditions, and then prioritize these scenarios/metrics in interpreting the results of this benchmark."* (Liang et al., 2022)

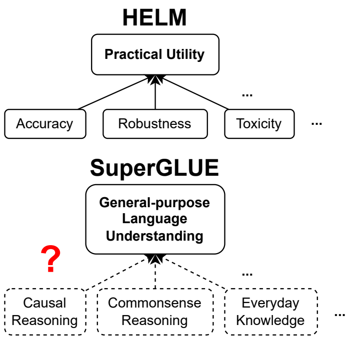

Figure 3: Different levels of capabilities and their connection, in HELM and SuperGLUE.

to measuring GLU(see Figure 3). When describing the selected datasets, the authors mention constructs like "causal reasoning,"[7] which could be seen as a sub-capability under GLU. However, these sub-capabilities are not defined, and their connection to GLU is left implied. By contrast, the choices and definitions of capabilities are much clearer in HELM. The benchmark aims to provide insights about LMs' "practical utility," and the seven capabilities of interest (e.g., "accuracy," "calibration") are selected as they reflect what *"it mean[s] for a system to be useful"* (Liang et al., 2022, p.27). The lack of clarity about the capabilities of interest makes it difficult to analyze whether a benchmark properly operationalizes them.

Capability module: The capabilities the benchmarks are purportedly measuring are often poorly and/or inconsistently conceptualized. The ECBD framework requires benchmark designers not only to say what they want to measure but also to justify why they want it. This helps us foreground inconsistencies in how these capabilities are defined and justified. For example, some of the analyzed benchmarks collapse the constructs they are designed to measure with the measurement of those constructs. Specifically, HELM describes e.g., "accuracy" (the construct) as an *"umbrella term for the standard accuracy-like metric"*(Liang et al., 2022, p.29) (possible measurements of the construct). This makes it difficult to even know what capability the resulting measurements actually measure. Furthermore, HELM also conceptualizes constructs like "fairness," "bias," and "toxicity" as measurable without requiring *''knowledge about the broader social context."*(Liang et al., 2022, p.28) We know, however, from prior work that more often than not, such constructs depend on the context in which they are applied (Blodgett et al., 2020). While such inconsistencies are not necessarily problematic, they can give rise to validity concerns if the benchmark's conceptualizations are not well-justified.

Content module: For benchmarks re-using data, we found little justification connecting the data to the capability of interest. When a benchmark re-uses pre-existing data, this data may not be originally designed to capture the capabilities of interest to this benchmark. By requiring benchmark designers to justify their choice of data, ECBD's content module helps highlight potential disconnect between the capability of interest and the re-used data. For instance, the BoolQ dataset was re-purposed by HELM to measure *(social) bias* amongst other capabilities. Since this dataset was not designed to elicit evidence about *bias*, ECBD requires HELM to justify (and validate) the re-use of this data to capture this capability. We found no such justification (nor validation), which raises doubts about whether the resulting *bias* measurement is meaningful.

Adaptation module: HELM gives great importance to its adaptation methods, while BoolQ and SuperGLUE do not prescribe any adaptation methods. ECBD's adaptation module draws attention to the suitability of adaptation methods, but only HELM prescribed an adaptation strategy: few-shot prompting with 5 in-context examples. Once chosen for a given dataset, these examples and the prompt template (e.g., instructions) stay fixed across across all test examples from that dataset, as well as across all evaluated models. By contrast, BoolQ and SuperGLUE do not specify how evaluated models/systems need to be adapted. As benchmark users are free to decide for themselves what methods to employ, it might become impossible to meaningfully interpret benchmark results when users adopt different adaptation methods for the same benchmark.

Assembly module: Benchmarks tend to overlook describing their assembly methods. ECBD, through the assembly module, emphasizes that the assembly methods are design choices that benchmark designers need to carefully consider. We find, however, that these choices—and the role they play in benchmarking—is largely overlooked. The authors of BoolQ only briefly mention that examples are split into training, development and test sets, without specifying how examples are selected to be part of the test set. For SuperGLUE, the

[7] *"COPA (Choice of Plausible Alternatives, Roemmele et al. (2011)) is a causal reasoning task [...]."* (Wang et al., 2019)

train/dev/test splits are in most cases already available from re-used datasets. For HELM, a maximum of 1,000 test examples per dataset are selected for evaluation, but we find no description about the exact selection process. This lack of attention to assembly methods could hinder benchmark designers from considering alternative methods (e.g., selecting examples based on their difficulty) and reflecting about trade-offs between benchmark quality and resource constraints.

Evidence module: The choice of evaluation methods is often justified by their adoption in prior work. All three benchmarks use automatic metrics to extract evidence, such as exact-match for classification tasks and ROUGE-2 for summarization. These metric scores are then accumulated through aggregation functions like F1-score and average. ECBD's evidence module requires benchmark designers to justify these choices, particularly with respect to the role they play in extracting and accumulating capability evidence. However, we find that existing justifications often do not focus on whether or how these methods capture the capabilities of interest. Instead, they are justified through brief mentions of the chosen metrics being "standard" or "default" for a certain task (Liang et al., 2022, p.127-137), or of the benchmark designers "follow[ing] prior work" (Wang et al., 2019, p.5-6) when choosing metrics or aggregation functions.

We encourage benchmark designers to more carefully consider their choices in the evidence module, including questioning methodology in prior work, so as not to risk perpetuating the use of currently popular yet potentially unsuitable methodology. Even where methods may be well-justified in prior work, they may not always be well-suited to other contexts (e.g., with differently defined capabilities under measurement), and their appropriateness to such new contexts should always be justified.

Evidence module: Even when new evaluation methods are introduced, we still find little justification for how the methods capture the capabilities of interest. For example, HELM introduces new automatic metrics to measure "(social) bias" through demographic representation. The metric first counts occurrences of words related to each considered demographic group (e.g., "gomez," "martinez," for the group "Hispanic") in model outputs. It then compares the word counts to the uniform distribution (i.e., where every demographic group is equally represented). The design decisions, such as the demographic groups under consideration and their corresponding word lists, are well-described. However, we found little justification for them. Why does the benchmark use the demographic groups "White," "Hispanic," and "Asian" to measure racial bias? Why is the uniform distribution a suitable reference distribution? Under ECBD, HELM would need to justify how these design decisions enable the new metric to capture "(social) bias."

The benchmarks rarely gather validity evidence to support their design decisions. All modules in the ECBD framework require collecting validity evidence. This step is either completely ignored, or acknowledged but left to future work. We encourage benchmark designers to search for and consider validity evidence that may already exist, and plan future experiments to gather necessary validity evidence. This step could require efforts from other researchers, benchmark users, etc. Proper incentives from the community could encourage future efforts on gathering validity evidence and on examining how to integrate this evidence into the use of existing benchmarks (e.g., how to use a benchmark that includes a metric which is found to be unsuitable?).

5 Conclusion

To guide benchmark creation and analysis, we take inspiration from the evidence-centered design framework from the field of educational testing to propose ECBD (Evidence-Centered Benchmark Design). Our framework formalizes the benchmark design process into five modules that each play a critical role in gathering reliable and valid capability evidence—i.e., evidence necessary to support the benchmark's measurement. We demonstrated its utility by analyzing BoolQ, SuperGLUE, and HELM, finding many common practices. For example, the benchmarks we analyzed tend to focus more on describing design choices (e.g., which dataset/metric is used), and less on justifying them and their role in the benchmark. Gathering validity evidence is also rare.

Future directions include analysis of our framework's utility in guiding the creation of benchmarks. As ECBD does not constrain the model inputs and outputs to be textual, we also see it to be applicable or adaptable to multi-modal NLP benchmarks, to to other areas in ML and AI.

Limitations

Findings from our case studies are limited by the choice of analyzed benchmarks: BoolQ, SuperGLUE, and HELM. Although these three benchmarks share many differences, they do not cover the wide space of possibilities in benchmark design. We have not analyzed, for instance, dynamic benchmarks that create test examples instead of relying on existing data (Kiela et al., 2021).

Furthermore, our analysis relied only on the papers introducing each of the three benchmarks, namely the work of Clark et al. (2019), of Wang et al. (2019), and of Liang et al. (2022). We have not used other sources of information on the benchmarks, such as their official websites and code repositories, which could limit our analysis. On the other hand, only relying on the papers allows us to examine the authors' reporting practice: what design choices do they prioritize given the limited space of an academic paper?

Finally, the case studies are subject to our reading. We could have missed or misinterpreted passages from the analyzed papers. Such mistakes in the completed worksheets could then impact our findings.

Ethical Considerations

NLP benchmarks not only influence the development and use of specific NLP systems, but could also shape the field when widely adopted by practitioners. As a result, well-documented and more valid benchmarks run less risk of misguiding benchmark users and stakeholders of evaluated systems—potentially avoiding the costs of optimizing systems towards the wrong goal, deploying systems with undetected issues and causing harms to system users, etc.

By proposing a more principled way of designing and analyzing NLP benchmarks, we hope to encourage the construction of well-documented and more valid benchmarks. However, our work could potentially have the unintended, opposite impact of discouraging future work in benchmark design. Although we believe that the benefits of following ECBD outweigh its costs, extensive documentation in following ECBD, as well as conducting experiments to gather validity evidence, could be expensive and time-consuming.

References

Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Tanti Wijaya. 2022. Challenges in measuring bias via open-ended language generation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 76–76, Seattle, Washington. Association for Computational Linguistics.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*.

National Council on Measurement in Education American Educational Research Association, American Psychological Association. 2014. *Standards for educational and psychological testing*. American Educational Research Association, Lanham, MD.

Deborah L. Bandalos. 2018. *Measurement theory and applications for the social sciences*. The Guilford Press.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings* of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. FairPrism: Evaluating fairness-related harms in text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6231–6251, Toronto, Canada. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez-Beltrachini, Leonardo F. R. Ribeiro, Lewis Tunstall, Li Zhang, Mahima Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. Gemv2: Multilingual nlg benchmarking in a single line of code.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring <mask>: evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Q. Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159.

Robert J Mislevy. 2003. On the Structure of Educational Assessments.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan

Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Kiri L. Wagstaff. 2012. Machine learning that matters. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1851–1856, Madison, WI, USA. Omnipress.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

A Worksheet Template

Introduction

Evidence-Centered Benchmark Design (ECBD) is a framework that formalizes the benchmark design process. It requires first specifying the **intended use** of the benchmark (including specifying the objects of evaluation). The process is then broken down into five modules:

i) **Capability module**: capabilities that the benchmark aims to measure.
ii) **Content module**: pool of test examples that draw out responses from the objects.
iii) **Adaptation module**: adapting or instructing the objects to complete the tasks.
iv) **Assembly module**: selecting from the pool of test examples to build the set used for evaluation.
v) **Evidence module**: extracting and accumulating evidence about the capabilities of interest from responses produced by the objects.

This worksheet provides guidance on how to create a new benchmark or analyze an existing benchmark following ECBD. It can be completed from different perspectives: as the creator of a new benchmark, as the custodian or the user of an existing benchmark, or as a third-party analyzing benchmarks, etc. Each module contains three questions:

- **Describe**: What design decisions did the benchmark creators make for this module?
- **Justify**: Why did the benchmark creators make these decisions? This involves forming a hypothesis that the decisions allow the module to accomplish its role in the process of gathering necessary capability evidence.
- **Support**: What validity evidence do the benchmark creators have to support the above hypothesis? In other words, what shows that the module indeed accomplishes its role?

This worksheet is not a checklist, and it is not required to answer each question perfectly. These questions are meant to encourage reflection and validation of benchmark design decisions, as well as to guide benchmark documentation.

Benchmark Name and Reference(s)

The references are the source of information used to complete this worksheet. For example, a third-party analyzing an existing benchmark may choose to use the academic publication introducing said benchmark as their source of information. Other

sources of information could be blog posts, official websites, or code repositories accompanying the benchmark.
[ANSWER HERE]

Who is filing the worksheet?

From what perspective is this worksheet completed? In other words, what is the relation between the person(s) completing this worksheet and the benchmark that is the focus of this worksheet?
[ANSWER HERE]

A.1 Intended Use

Q1 - Who/What are the intended objects of evaluation? Elaboration on the objects of evaluation (e.g., their assumed capabilities, demographic information for human objects of evaluation, etc.) helps us better understand whether the benchmark is suitable for all intended objects of evaluation.
[ANSWER HERE]

Q2 - What is the intended use of the benchmark? Who are the intended users of the benchmark? Benchmark results aim to provide insights about the objects of evaluation: how are users meant to use these insights?
[ANSWER HERE]

A.2 Capability Module

The capability module specifies the capabilities that the benchmark aims to evaluate. The term "capability" refers to a construct (e.g., quality criteria, skill, etc.) that the objects of evaluation are thought to exhibit or possess. Capabilities often cannot be directly observed or directly measured, thus requiring the benchmark to indirectly measure them by gathering necessary evidence about said capabilities.

Q3 - DESCRIBE: i) What are the capabilities of interest? ii) How is each one defined, and under what context is each one defined?
[ANSWER HERE]

Additional recommended questions to consider so to further clarify and contextualize the definitions (in benchmark analysis: as presented by the benchmark):

- How does the definition used by the benchmark differ from other existing definitions of this capability?
  [ANSWER HERE]
- How does this capability differ from other similarly defined capabilities?
  [ANSWER HERE]

Q4 - JUSTIFY: How are the capabilities of interest connected to the intended use of the benchmark (specified in Q2)? Are the capabilities theoretically attainable by the objects to be evaluated? Explain the interest in measuring the capabilities in Q3 and question whether it may be impossible for the objects of evaluation to have said capabilities.
[ANSWER HERE]

Q5 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice and definition of capabilities of interest?
[ANSWER HERE]

A.3 Content Module

The content module specifies test examples that the benchmark could require objects of evaluation to perform or to respond to. The examples should elicit evidence about some capability of interest, so that said capability evidence can be later extracted from the responses and aggregated to produce a measurement of said capability.

Q6 - DESCRIBE: i) Characterize the examples. Most often, NLP evaluation relies on input data, so this step could involve describing the data that is available to the benchmark to use, how the data is obtained, etc. **ii) Which capabilities of interest does each example aim to capture?** Each example can aim to capture one or several capabilities amongst those listed in Q3.
[ANSWER HERE]

Q7 - JUSTIFY: How does each example elicit evidence about its target capabilities? Justify via the characteristics of the examples (Q6).
[ANSWER HERE]

Q8 - SUPPORT: What evidence do the benchmark creators offer to support content validity of the test examples? In other words, we question whether the test examples captures capabilities of interest. Content validity is often based on analysis by external experts or benchmark users.
[ANSWER HERE]

A.4 Adaptation Module

When evaluating humans, the benchmark might instruct them to perform a task by providing instructions, training exercises, demonstrations, etc. When evaluating models/systems, there are also myriad methods that i) modify the models/systems (e.g., fine-tuning), or ii) format or add onto the input (e.g., adding examples in few-shot prompting). These adaptation methods should be chosen carefully so as to not confound evaluation results.

Q9 - DESCRIBE: Given an input, how are the objects of evaluation adapted or instructed to provide the output?
[ANSWER HERE]

Q10 - JUSTIFY: Elaborate on the suitability of the adaptation methods for all intended objects of evaluation.
[ANSWER HERE]

Q11 - SUPPORT: What validity evidence do benchmark designers offer that supports the choice of the adaptation methods?
[ANSWER HERE]

A.5 Assembly Module

Examples specified by the content module are what the benchmark could use. The assembly module concerns what examples from that pool will actually be used by the benchmark for evaluation, and whether this set allows the benchmark to gather sufficient evidence.

Q12 - DESCRIBE: How many examples are chosen to assemble the subset used for evaluation? What factors inform this selection?
[ANSWER HERE]

Q13 - JUSTIFY: How does the described assembly method ensure that the produced subset elicits sufficient evidence for all capabilities of interest?
[ANSWER HERE]

Q14 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice of assembly methods?
[ANSWER HERE]

A.6 Evidence Module

A.6.1 Evidence Extraction Component

In response to each presented test example, objects of evaluation produce observable behaviors (referred to as “responses”) which are captured by the benchmark. From these responses, the benchmark extracts evidence about capabilities of interest that said test example targets (referred to as “salient evidence”).

Q15 - DESCRIBE: For each test example, i) What responses are captured and used for evidence extraction? When evaluating humans, many types of responses can be captured: selection in multiple-choice questions, long-form answers, response time, etc. Similarly, the benchmark can use the generated text (decoded in a certain way), token probabilities, running time, etc. **ii) How is evidence extracted and represented?**
[ANSWER HERE]

Q16 - JUSTIFY: How does the extracted evidence capture the capabilities of interest?
[ANSWER HERE]

Q17 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice of evidence extraction method?
[ANSWER HERE]

A.6.2 Evidence Accumulation Component

Q18 - DESCRIBE: How is the evidence accumulated to draw insights about the objects of evaluation in terms of capabilities of interest?
[ANSWER HERE]

Q19 - JUSTIFY: How does the method of accumulating evidence capture capabilities of interest?
[ANSWER HERE]

Q20 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice of evidence accumulation method?
[ANSWER HERE]

B Glossary

We compile terminology used in the present paper and in the ECBD worksheet in Table 1.

Term	Meaning
Objects of evaluation	Models, systems, people, etc. that are to be evaluated.
Capability	Quality criteria, ability, skill, etc. that characterizes the objects of evaluation. They are very often not observable nor directly measurable.
Capability evidence	Evidence indicating whether or to what degree an object of evaluation has the capability of interest. For example, a language model (object of evaluation) detecting the grammatical error in "their going to the mall." can be a piece of evidence supporting the belief that the model has grammatical knowledge (capability of interest).
Benchmarking (verb); a benchmark (noun)	We view benchmarking as a process of gathering capability evidence from the objects of evaluation about the capabilities of interest. A benchmark is a collection of measurement instruments that supports the above process.
Benchmark results	The final product of benchmarking, often in the form of numerical scores (e.g., ratio), rankings, or categorization (e.g., detecting that an object of evaluation is "biased"). The results inform benchmark users about the objects of evaluation, about to whether or to what degree the object has the capabilities of interest.
Validity Evidence	Evidence supporting whether the benchmark results can be interpreted as it is originally intended to be interpreted, whether the benchmark can be used as it is originally intended to be used. In other words, it is evidence supporting that the capability evidence gathered is actually meaningful with respect to the intended uses of the benchmark. Validity evidence can be theoretical or empirical/
Validity; validation	Validity is the degree to which all the accumulated validity evidence supports the intended interpretation of benchmark results for the intended use of the benchmark. Validation is thus the process of accumulating validity evidence
Test example	A single evaluation instance of the benchmark that objects of evaluation can be asked to perform or respond to in order to obtain outputs or behaviours from them.
Response	Outputs or behaviours from the objects of evaluation in response to a test example presented to them. These are expected to be observable. For example, a matrix of token probabilities can be a response from a language model. The decoded text that the model generated can also be a response. What response to capture is a benchmark design decision.
Context (in the capability module)	Where and how the objects of evaluation are intended to be used or intended to operate under. Context can involve the types of model/system users, other stakeholders, the domain of application, the linguistic phenomena the systems are meant to represent, etc. The definition of capabilities can greatly vary depending on context (e.g., informativeness of some texts varies for expert vs. non-expert readers)

Table 1: Glossary